Component Evolution in General Random Intersection Graphs

Milan Bradonjić, Aric Hagberg, Nicolas W. Hengartner, Allon G. Percus

Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, milan@lanl.gov,
Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, hagberg@lanl.gov,
Information Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, nickh@lanl.gov,
School of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, USA, allon.percus@cgu.edu.

Abstract. Random intersection graphs (RIGs) are an important random structure with algorithmic applications in social networks, epidemic networks, blog readership, and wireless sensor networks. RIGs can be interpreted as a model for large randomly formed non-metric data sets. We analyze the component evolution in general RIGs, giving conditions on the existence and uniqueness of the giant component. Our techniques generalize existing methods for analysis of component evolution: we analyze survival and extinction properties of a dependent, inhomogeneous Galton-Watson branching process on general RIGs. Our analysis relies on bounding the branching processes and inherits the fundamental concepts of the study of component evolution in Erdős-Rényi graphs. The major challenge comes from the underlying structure of RIGs, which involves both a set of nodes and a set of attributes, with different probabilities associated with each attribute.

Keywords: General random intersection graphs, random graphs, branching processes, giant component, stochastic processes in relation with random discrete structures.

1 Introduction

Bipartite graphs, consisting of two sets of nodes with edges only connecting nodes in opposite sets, are a natural representation for many algorithmic problems on networks. Social networks can often be cast as bipartite graphs built from sets of individuals connected to sets of attributes, such as membership of a club or organization, work colleagues, or fans of the same sports team. A well-known example is a collaboration graph, where the two sets might be scientists and research papers, or actors and movies [27, 18]. Simulations of epidemic spread in human populations are often performed on networks constructed from bipartite graphs of people and the locations they visit during a typical day [11]. Bipartite structure is hardly limited to social networks. The relation between nodes and keys in secure wireless communication, for examples, forms a bipartite network [6]. Factor graphs have become a standard representation for

constraint satisfaction problems such as k-SAT and graph coloring. In general, bipartite graphs are well suited to problems of classifying objects, where each object has a set of properties [10]. However, modeling such networks remains a challenge. The well-studied Erdős-Rényi model, $G_{n,p}$, successfully used for average-case analysis of algorithm performance, does not satisfactorily represent many randomly formed social or collaboration networks. $G_{n,p}$ does not capture the typical scale-free degree distribution of many real-world networks [3]. More realistic degree distributions can be achieved by the configuration model [20] or expected degree model [7], but even those fail to capture common properties of social networks such as the high number of triangles (or cliques) and strong degree-degree correlation [19, 1].

A straightforward way of remedying these problems is to characterize each of the bipartite sets separately. One step in this direction is an extension of the configuration model that specifies degrees in both sets [14]. We study the related approach of random intersection graphs (RIG), first introduced in [26, 16]. Any undirected graph can be represented as an intersection graph [9]. The simplest version is the "uniform" RIG, $G(n, m, p)$, containing a set of n nodes and a set of m attributes, where any given node-attribute pair contains an edge with a fixed probability p, independently of other pairs. Two nodes in the graph are taken to be connected if and only if they are both connected to at least one common element in the attribute set. In our work, we study the more general RIG, $G(n, m, \boldsymbol{p})$ [22, 21], where the node-attribute edge probabilities are not given by a uniform value p but rather by a set $\boldsymbol{p} = \{p_w \mid w \in W\}$. A node is attached to the attribute w, with probability p_w.[1] This general model has only recently been developed and only a few results have been obtained, such as expander properties, cover time, and the existence and efficient construction of large independent sets [22, 21, 23].

In this paper, we generalize results that have previously been obtained for the uniform RIG [4, 6], analyzing the evolution of components in general RIGs and obtaining conditions for the existence and uniqueness of the giant component. Our main contribution is a generalization of the branching process used for analyzing $G_{n,p}$ [2]. By considering an auxiliary process that is stochastically equivalent, we bound the stopping time for the branching process on general RIGs, yielding bounds on the sizes of graph components. The major challenge comes from the underlying structure of RIGs, which involves both the set of nodes and the set of attributes, as well as the set of different probabilities $\boldsymbol{p} = \{p_w \mid w \in W\}$. Our approach requires us to keep track of the history of the branching process, which is directly dictated by this structure.

2 Model and previous work

In this paper, we will consider the general intersection graph $G(n, m, \boldsymbol{p})$, introduced in [22, 21], with a set of probabilities $\boldsymbol{p} = \{p_w \mid w \in W\}$, where $p_w \in (0, 1)$. We now formally define the model.

Model. Given a set of nodes $V = \{1, 2, \ldots, n\}$, attributes $W = \{1, 2, \ldots, m\}$, and probabilities $\boldsymbol{p} = \{p_w \mid w \in W\}$, for all $(v, w) \in V \times W$, define the i.i.d. indicator

[1] Note that the p_w do not generally sum up to 1. Furthermore, we can eliminate the trivial cases of $p_w = 0$ and $p_w = 1$, corresponding to the absence of attribute w and to a complete graph.

random variables

$$I_{v,w} \sim \text{Bernoulli}(p_w). \tag{1}$$

Every node $v \in V$ is assigned a random set of attributes $W(v) \subseteq W$

$$W(v) = \{w \in W \mid I_{v,w} = 1\}. \tag{2}$$

This is illustrated schematically in Fig. 1.

A set of edges $E \in V \times V$ is then defined, such that for two different nodes $v_i, v_j \in V$, $\{v_i, v_j\} \in E$ iff

$$|W(v_i) \cap W(v_j)| \geq s \tag{3}$$

for a given integer $s \geq 1$. Thus, two nodes are adjacent if and only if they have at least s attributes in common. One limitation of our analysis is that for simplicity, we fix $s = 1$.

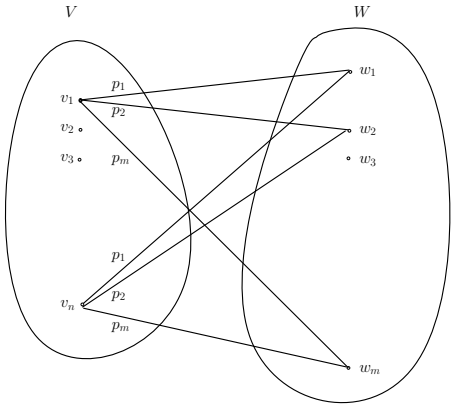

Fig. 1. Random intersection graph. V is set of nodes and W is set of attributes. A particular attribute w_i is associated with every node independently at random with probability p_i.

Our model generalizes the uniform model $G(n, m, p)$, studied in [4, 6], where all p_w take on the same value p. Different generalizations and special cases have been studied in [13, 15, 17, 8].

To complete the picture of previous work, in [8] it was shown that when $n = m$, a set of probabilities $\boldsymbol{p} = \{p_w \mid w \in W\}$ can be chosen to tune the degree and clustering coefficient of the graph.

3 Mathematical preliminaries

In this paper, we analyze the component evolution of the general RIG structure. As we have already mentioned, the major challenge comes from the underlying structure of RIGs, which involves both a set of nodes and a set of attributes, as well as a set of different probabilities $\boldsymbol{p} = \{p_w \mid w \in W\}$.

Moreover, the edges in RIG are not independent. Hence, a RIG cannot be treated as an Erdős-Rényi random graph $G_{n,\hat{p}}$, with the edge probability $\hat{p} = 1 - \prod_{w \in W}(1 - p_w^2)$. However, in [12], the authors provide the comparison among $G_{n,\hat{p}}$ and $G(n, m, p)$, showing that for $m = n^\alpha$ and $\alpha > 6$, these two classes of graphs have asymptotically the same properties. In [25], Rybarczyk has recently shown the equivalence of sharp threshold functions among $G_{n,\hat{p}}$ and $G(n, m, p)$, when $m \geq n^3$. In this work, we do not impose any constraints among n and m, and we develop methods for the analysis of branching processes on RIGs, since the existing methods for the analysis of branching processes on $G_{n,p}$ do not apply.

We now briefly state the edge dependence. Consider three distinct nodes v_i, v_j, v_k from V, and let "$\leftrightarrow$" denote adjacency, so that $v_i \leftrightarrow v_j$ iff $|W(v_i) \cap W(v_j)| \geq 1$. Conditional on the set $W(v_k)$, by the definition (2), the sets $W(v_i) \cap W(v_k)$ and $W(v_j) \cap W(v_k)$ are mutually independent, which implies conditional independence of the events $\{v_i \leftrightarrow v_k \mid W(v_k)\}, \{v_j \leftrightarrow v_k \mid W(v_k)\}$, that is,

$$\mathbb{P}[v_i \leftrightarrow v_k, v_j \leftrightarrow v_k \mid W(v_k)] = \mathbb{P}[v_i \leftrightarrow v_k \mid W(v_k)]\,\mathbb{P}[v_j \leftrightarrow v_k \mid W(v_k)]. \tag{4}$$

However, the latter does not imply independence of the events $\{v_i \leftrightarrow v_k\}$ and $\{v_j \leftrightarrow v_k\}$ since in general

$$\begin{aligned}\mathbb{P}[v_i \leftrightarrow v_k, v_j \leftrightarrow v_k] &= \mathbb{E}\big[\mathbb{P}[v_i \leftrightarrow v_k, v_j \leftrightarrow v_k \mid W(v_k)]\big] \\ &= \mathbb{E}\big[\mathbb{P}[v_i \leftrightarrow v_k \mid W(v_k)]\,\mathbb{P}[v_j \leftrightarrow v_k \mid W(v_k)]\big] \\ &\neq \mathbb{P}[v_i \leftrightarrow v_k]\,\mathbb{P}[v_j \leftrightarrow v_k]. \end{aligned} \tag{5}$$

Furthermore, the conditional pairwise independence (4) does not extend to three or more nodes. Indeed, conditionally on the set $W(v_k)$, the sets $W(v_i) \cap W(v_j), W(v_i) \cap W(v_k)$, and $W(v_j) \cap W(v_k)$ are not mutually independent, and hence neither are the events $\{v_i \leftrightarrow v_j\}, \{v_i \leftrightarrow v_k\}$, and $\{v_j \leftrightarrow v_k\}$, that is,

$$\begin{aligned}\mathbb{P}[v_i \leftrightarrow v_j, v_i \leftrightarrow v_k, v_j \leftrightarrow v_k \mid W(v_k)] \neq \mathbb{P}[v_i \leftrightarrow v_j \mid W(v_k)] \\ \times \mathbb{P}[v_i \leftrightarrow v_k \mid W(v_k)]\,\mathbb{P}[v_j \leftrightarrow v_k \mid W(v_k)].\end{aligned}$$

4 Auxiliary process on general random intersection graphs

Our analysis for the emergence of a giant component is inspired by the process described in [2], which measures the size of a component by counting the number of steps until a breadth-first search terminates. The difficulty in using this approach to analyze the evolution of the stochastic process defined by equations (1), (2), and (3) resides in the fact that we need, at least in principle, to keep track of the temporal evolution of the sets of nodes and attributes being explored. This results in a process that is not Markovian.

Therefore, we instead construct an auxiliary process that is simpler to analyze but whose stopping time is, in distribution, identical to that of the breadth-first search. The process is algorithmically defined as follows.

Auxiliary Process. Start from an arbitrary node $v_0 \in V$. Denote by V_t the cumulative set of nodes *visited* by time t, which we initialize to $V_0 = \{v_0\}$. Denote the cumulative set of all attributes [6] associated with the set V_t by

$$W_t = \bigcup_{\tau=0}^{t} W(v_\tau). \tag{6}$$

Now consider the set of nodes adjacent to V_t but not yet visited:

$$\Big\{v \in V \setminus V_t : W(v) \cap W_t \neq \emptyset\Big\}. \tag{7}$$

Following [2], we call this the set of *alive* nodes at time t. Unlike in [2], however, we do not keep track of the actual list of alive nodes, but only the size of the set, which we denote by the random variable

$$Y_t = \left| \left\{ v \in V \setminus V_t : W(v) \cap W_t \neq \emptyset \right\} \right|,$$

for $t \geq 1$. We also define $Y_0 = |V_0|$, that is, $Y_0 = 1$. The process evolves as follows: for $t \geq 1$, pick a node v_t uniformly at random from the set $V \setminus V_{t-1}$ and update the set of visited nodes $V_t = V_{t-1} \cup \{v_t\} = \{v_0, \ldots, v_t\}$. Then update the set of alive nodes and Y_t. The process terminates when Y_t reaches 0.

To understand why this auxiliary process is useful, notice that due to the independence of the random variables $I_{v,w}$, at step t every node in $V \setminus V_t$ is equally likely to belong to the set (7) of alive nodes. Consequently, picking the next node v_{t+1} uniformly from $V \setminus V_t$ is the same random process as picking v_{t+1} uniformly from the set of alive nodes (as in [2]), *conditional* on the history of the attribute sets uncovered up through time t:

$$\mathcal{H}_t = \{W(v_0), W(v_1), \ldots, W(v_t)\}. \tag{8}$$

In the latter process, the stopping time

$$T(v_0) = \inf\{t \geq 0 : Y_t = 0\} \tag{9}$$

would simply be equal to $|C(v_0) - 1|$, where $|C(v_0)$ is the size of the component containing v_0 [2].Thus in the auxiliary process, since it is stochastically equivalent,

$$T(v_0) \stackrel{d}{=} |C(v_0)| - 1. \tag{10}$$

4.1 Process description in terms of random variable Y_t

Let us characterize the process $\{Y_t\}_{t \geq 0}$ in terms of the number Z_t of newly discovered neighbors of V_t:

$$Z_t = Y_t - Y_{t-1} + 1, \tag{11}$$

where the term +1 reflects the fact that v_t is discovered at time step t, but it is not counted in Y_t because it has been visited. For nodes that are neither visited nor alive, the events of their becoming alive at time t are conditionally independent given the history $\mathcal{H}_t$, since each event involves a different subsets of the indicator random variables $I_{v,w}$. $W(v_t)$ and W_{t-1} are mutually independent, hence the conditional probability that such a node u becomes alive at time t is

$$\begin{aligned}
r_t &= \mathbb{P}[u \leftrightarrow v_t, u \not\leftrightarrow v_{t-1}, u \not\leftrightarrow v_{t-2}, \ldots, u \not\leftrightarrow v_0 | \mathcal{H}_t] \\
&= \mathbb{P}[W(u) \cap W(v_t) \neq \emptyset, W(u) \cap W_{t-1} = \emptyset | \mathcal{H}_t] \\
&= \mathbb{P}[W(u) \cap W(v_t) \neq \emptyset, W(u) \cap W_{t-1} = \emptyset | W(v_t), W_{t-1}] \\
&= \mathbb{P}[W(u) \cap W(v_t) \neq \emptyset | W(v_t)] \, \mathbb{P}[W(u) \cap W_{t-1} = \emptyset | W_{t-1}] \\
&= \Big(1 - \prod_{\alpha \in W(v_t)} (1 - p_\alpha)\Big) \prod_{\beta \in W_{t-1}} (1 - p_\beta).
\end{aligned}$$

The last expression can be rewritten as

$$\begin{aligned} r_t &= \prod_{\beta \in W_{t-1}} q_\beta - \prod_{\alpha \in W_t} q_\alpha \\ &= \phi_{t-1} - \phi_t, \end{aligned} \tag{12}$$

where we set $q_w = 1 - p_w$ for all $w \in W$ and $\phi_t = \prod_{\alpha \in W_t} q_\alpha$, and use the convention $W_{-1} = \emptyset$ and $\phi_{-1} = 1$.

Observe that the probability (12) does not depend on u. Hence the number of new alive nodes at time t is, conditional on the history $\mathcal{H}_t$, a Binomial distributed random variable with parameters r_t and $N_t = n - t - Y_t$:

$$Z_{t+1} | \mathcal{H}_t \sim \text{Bin}(N_t, r_t). \tag{13}$$

Now, by mathematical induction in t, it easily follows that for times $t \geq 1$ the number of alive nodes Y_t satisfies:

$$Y_t | \mathcal{H}_{t-1} \sim \text{Bin}\Big(n - 1, 1 - \prod_{\tau=0}^{t-1} (1 - r_\tau)\Big) - t + 1. \tag{14}$$

4.2 Expectation and variance of ϕ_t

The history $\mathcal{H}_t$ embodies the evolution of how the attributes are discovered over time. It is insightful to recast that history in terms of the discovery times Γ_w of each attribute in W. Given any sequence of nodes $v_0, v_1, v_2, \ldots$, the probability that a given attribute w is first discovered at time $t < n$ is

$$\begin{aligned} \mathbb{P}[\Gamma_w = t] &= \mathbb{P}[I_{v_t,w} = 1, I_{v_{t-1},w} = 0, \ldots, I_{v_0,w} = 0] \\ &= p_w (1 - p_w)^t. \end{aligned}$$

If an attribute w is not discovered by time $n - 1$, we set $\Gamma_w = \infty$ and note that

$$\mathbb{P}[\Gamma_w = \infty] = (1 - p_w)^n.$$

From the independence of the random variables $I_{v,w}$, it follows that the discovery times $\{\Gamma_w : w \in W\}$ are mutually independent. We now focus on describing the distribution of $\phi_t = \prod_{\alpha \in W_t} q_\alpha$. For $t \geq 0$, we have

$$\phi_t = \prod_{\alpha \in W_t} q_\alpha = \prod_{j=0}^{t} \prod_{\alpha \in W_j \setminus W_{j-1}} q_\alpha \stackrel{d}{=} \prod_{j=0}^{t} \prod_{w \in W} q_w^{\mathbb{I}(\Gamma_w = j)} = \prod_{w \in W} q_w^{\mathbb{I}(\Gamma_w \leq t)}. \tag{15}$$

$$\mathbb{E}[\phi_t] = \prod_{w \in W} \Big(1 - (1 - q_w)(1 - q_w^{t+1})\Big). \tag{16}$$

The concentration of ϕ_0 will be crucial for the analysis of the supercritical regime. Hence, we provide $\mathbb{E}[\phi_0]$ and $\mathbb{E}[\phi_0^2]$ here. In Subsection 5.2, we will assume that $p_w = p(\log n / n)$. Under this condition, it follows from (16) that

$$\mathbb{E}[\phi_0] = \prod_{w \in W} (1 - p_w^2) = 1 - \sum_{w \in W} p_w^2 + o(\sum_{w \in W} p_w^2). \tag{17}$$

Moreover, under the same condition, it follows from (15) that

$$\mathbb{E}[\phi_0^2] = \mathbb{E}[\prod_{w\in W} q_w^{2\mathbb{I}(\Gamma_w\le 0)}] = \prod_{w\in W}\Big(1-(1-q_w^2)\mathbb{P}[\Gamma_w=0]\Big) = \prod_{w\in W}\Big(1-(1-q_w^2)p_w\Big)$$
$$= \prod_{w\in W}\Big(1-2p_w^2+p_w^3\Big) = 1-2\sum_{w\in W}p_w^2 + o(\sum_{w\in W}p_w^2). \tag{18}$$

5 Giant component

With the process $\{Y_t\}_{t\ge 0}$ defined in the previous section, we analyze both the sub-critical and supercritical regime of a general random intersection graph by adapting the percolation-based techniques used to analyze Erdős-Rényi random graphs [2]. The technical difficulty in analyzing that stopping time rests in the fact that the distribution of Y_t depends on the history of the process, dictated by the structure of the general RIG. In the next two subsections, we will give conditions for the absence as well as for the existence and uniqueness of the giant component, in general RIGs.

5.1 Subcritical regime

Theorem 1. *Suppose that*

$$p_w = O(1/n) \text{ for all } w \quad \text{and} \quad \sum_{w\in W} p_w^3 = O(1/n^2).$$

For any positive constant $c<1$, if $\sum_{w\in W} p_w^2 = c/n$, then ***whp***[2] *all components in a general random intersection graph $G(n,m,\boldsymbol{p})$ are of order $O(\log n)$.*

Proof. We generalize the techniques used in the proof for the sub-critical case in $G_{n,p}$ presented in [2]. Let $T(v_0)$ be the stopping time defined in (9), for the process starting at node v_0 and recall that $T(v_0) \stackrel{d}{=} |C(v_0)|$. We will bound the size of the largest component, and prove that under the conditions of the theorem, all components are of order $O(\log n)$, **whp**.

For all $t\ge 0$,

$$\mathbb{P}[T(v_0)>t] = \mathbb{E}\big[\mathbb{P}[T(v_0)>t \mid \mathcal{H}_t]\big] \le \mathbb{E}\big[\mathbb{P}[Y_t>0 \mid \mathcal{H}_t]\big]$$
$$= \mathbb{E}\left[\mathbb{P}[\mathrm{Bin}(n-1, 1-\prod_{\tau=0}^{t-1}(1-r_\tau)) \ge t \mid \mathcal{H}_t]\right]. \tag{19}$$

Bounding from above, we have

$$1-\prod_{\tau=0}^{t-1}(1-r_\tau) \le \sum_{\tau=0}^{t-1} r_\tau = \sum_{\tau=0}^{t-1}(\phi_{\tau-1}-\phi_\tau) = 1-\phi_{t-1}, \tag{20}$$

[2] "With high probability," meaning with probability $1-o(1)$, as the number of nodes $n\to\infty$.

which can readily be shown by induction in t for $r_\tau \in [0,1]$. By using stochastic ordering of the Binomial distribution, both in n and in $\sum_{\tau=0}^{t-1} r_\tau$, and for any positive constant $\nu < 1$, which is to be specified later, it follows that

$$\begin{aligned}
&\mathbb{P}[T(v_0) > t \mid \mathcal{H}_t] \leq \mathbb{P}[\mathrm{Bin}(n, \sum_{\tau=0}^{t-1} r_\tau) \geq t \mid \mathcal{H}_t] \leq \mathbb{P}[\mathrm{Bin}(n, 1-\phi_{t-1}) \geq (1-\nu)t \mid \mathcal{H}_t] \\
&= \mathbb{P}[\mathrm{Bin}(n, 1-\phi_{t-1}) \geq t \mid 1-\phi_{t-1} < (1-\nu)t/n \cap \mathcal{H}_t]\mathbb{P}[1-\phi_{t-1} < (1-\nu)t/n \mid \mathcal{H}_t] \\
&\quad + \mathbb{P}[\mathrm{Bin}(n, 1-\phi_{t-1}) \geq t \mid 1-\phi_{t-1} \geq (1-\nu)t/n \cap \mathcal{H}_t]\mathbb{P}[1-\phi_{t-1} \geq (1-\nu)t/n \mid \mathcal{H}_t] \\
&\leq \mathbb{P}[\mathrm{Bin}(n, 1-\phi_{t-1}) \geq t \mid 1-\phi_{t-1} < (1-\nu)t/n \cap \mathcal{H}_t] \\
&\quad + \mathbb{P}[1-\phi_{t-1} \geq (1-\nu)t/n \mid \mathcal{H}_t]. \qquad (21)
\end{aligned}$$

Furthermore, using the fact that the event $\{1-\phi_{t-1} < (1-\nu)t/n\}$ is $\mathcal{H}_t$-measurable, together with the stochastic ordering of the binomial distribution, we obtain

$$\mathbb{P}[\mathrm{Bin}(n, 1-\phi_{t-1}) \geq t \mid 1-\phi_{t-1} < (1-\nu)t/n \cap \mathcal{H}_t] \leq \mathbb{P}[\mathrm{Bin}(n, (1-\nu)t/n) \geq t \mid \mathcal{H}_t].$$

Taking the expectation with respect to the history $\mathcal{H}_t$ in (21) yields

$$\mathbb{P}[T(v_0) > t] \leq \mathbb{P}[\mathrm{Bin}(n, (1-\nu)t/n) \geq t] + \mathbb{P}[1-\phi_{t-1} \geq (1-\nu)t/n].$$

For $t = K_0 \log n$, where K_0 is a constant large enough and independent on the initial node v_0, the Chernoff bound ensures that $\mathbb{P}[\mathrm{Bin}(n, (1-\nu)t/n) \geq t] = o(1/n)$, and

$$\begin{aligned}
&\mathbb{P}\{1-\phi_{t-1} \geq (1-\nu)t/n\} = \mathbb{P}\left\{\prod_{w \in W} q_w^{\mathbb{I}(\Gamma_w \leq t)} \leq 1 - \frac{(1-\nu)t}{n}\right\} \\
&= \mathbb{P}\left\{\sum_{w \in W} \log\left(\frac{1}{1-p_w}\right)\mathbb{I}(\Gamma_w \leq t) \geq -\log\left(1 - \frac{(1-\nu)t}{n}\right)\right\} \\
&\leq \mathbb{P}\left\{\sum_{w \in W} \log\left(\frac{1}{1-p_w}\right)\mathbb{I}(\Gamma_w \leq t) \geq \frac{(1-\nu)t}{n}\right\}. \qquad (22)
\end{aligned}$$

Define the auxiliary random variables $X_{t,w} = n\log(1/(1-p_w))\mathbb{I}(\Gamma_w \leq t)$, so that

$$\begin{aligned}
\mathbb{E}[X_{t,w}] &= n\log\Big(\frac{1}{1-p_w}\Big)(1-q_w^t) = n\Big(p_w + o(p_w)\Big)\Big(1-(1-p_w)^t\Big) \\
&= n\Big(p_w + o(p_w)\Big)\Big(tp_w + o(tp_w)\Big) = ntp_w^2 + o\big(ntp_w^2\big), \qquad (23)
\end{aligned}$$

which implies

$$\sum_{w \in W} \mathbb{E}[X_{t,w}] = nt \sum_{w \in W} p_w^2\big(1 + o(1)\big). \qquad (24)$$

Thus under the stated condition that

$$n \sum_{w \in W} p_w^2 = c < 1,$$

it follows that for some constant c', where $c < c' < 1$, and for sufficiently large n, $\sum_{w\in W}\mathbb{E}[X_{t,w}] \le c't$. In light of (22) and Bernstein's inequality [5],

$$\begin{aligned}\mathbb{P}\Big[1-\phi_{t-1} \ge \frac{(1-\nu)t}{n}\Big] &\le \mathbb{P}\Big[\sum_{w\in W} X_{t,w} \ge (1-\nu)t\Big] \\ &\le \mathbb{P}\Big[\sum_{w\in W}\big(X_{t,w}-\mathbb{E}[X_{t,w}]\big) \ge (1-\nu-c')t\Big] \\ &\le \exp\Big\{\frac{-\frac{3}{2}((1-\nu-c')t)^2}{3\sum_w \mathrm{Var}[X_{t,w}] + nt\max_w p_w(1+o(1))}\Big\} \qquad (25)\end{aligned}$$

Since

$$\begin{aligned}\mathbb{E}[X_{t,w}^2] &= \Big(n\log\Big(\frac{1}{1-p_w}\Big)\Big)^2(1-q_w^t) = n^2\Big(p_w+o(p_w)\Big)^2\Big(1-(1-p_w)^t\Big) \\ &= n^2\Big(p_w^2+o(p_w^2)\Big)\Big(tp_w+o(tp_w)\Big) = n^2tp_w^3+o\Big(n^2tp_w^3\Big), \qquad (26)\end{aligned}$$

it follows that for some large constant $K_1 > 0$

$$\sum_{w\in W}\mathrm{Var}[X_{t,w}] \le \sum_{w\in W}\mathbb{E}[X_{t,w}^2] = n^2t\sum_{w\in W}p_w^3 + o\Big(n^2t\sum_{w\in W}p_w^3\Big) \le K_1t.$$

Finally, the assumption of the theorem implies that there exists a constant $K_2 > 0$ such that

$$n\max_{w\in W}p_w \le K_2.$$

Substituting these bounds into (25) yields

$$\mathbb{P}[1-\phi_{t-1} \ge (1-\nu)t/n] \le \exp\left(-\frac{3(1-\nu-c')^2}{2(3K_1+K_2)}t\right).$$

Taking $\nu \in (0, 1-c')$ and $t = K_3\log n$ for some constant K_3 large enough and not depending on the initial node v_0, we conclude that $\mathbb{P}[1-\phi_{t-1} \ge (1-\nu)t/n] = o(n^{-1})$, which in turn implies that taking constant $K_4 = \max\{K_0, K_3\}$, ensures that

$$\mathbb{P}[T(v_0) > K_4\log n] = o(1/n)$$

for any initial node v_0. Finally, the union bound over the n possible starting values v_0 gives

$$\mathbb{P}[\max_{v_0\in V}T(v_0) > K_4\log n] \le no(n^{-1}) = o(1),$$

which implies that all connected components are of size $O(\log n)$, **whp**.

5.2 Supercritical regime

We now turn to the study of the supercritical regime in which $\lim_{n\to\infty} n\sum_{w\in W}p_w^2 = c > 1$.

Theorem 2. *Suppose that*

$$p_w = o\Big(\frac{\log n}{n}\Big) \text{ for all } w \quad \text{and} \quad \sum_{w\in W} p_w^3 = o\Big(\frac{\log n}{n^2}\Big).$$

For any constant $c > 1$, if $\sum_{w\in W} p_w^2 = c/n$, then ***whp*** *there exists a unique largest component in $G(n, m, \boldsymbol{p})$, of order $\Theta(n)$. Moreover, the size of the giant component is given by $n\zeta_c(1 + o(1))$, where ζ_c is the solution in $(0, 1)$ of the equation $1 - e^{-c\zeta} = \zeta$, while all other components are of size $O(\log n)$.*

Remark. The conditions on p_w and $\sum_w p_w^3$ are weaker than ones in the case of the subcritical regime.

Proof. We start by bounding $1 - \prod_{\tau=0}^{t-1}(1 - r_\tau)$. The upper bound $\sum_{\tau=0}^{t-1} r_\tau$ has already been established in (20). For the lower bound, we apply Jensen's inequality to the function $\log(1 - x)$ to get

$$\begin{aligned}\log \prod_{\tau=0}^{t-1}(1 - r_\tau) &= \sum_{\tau=0}^{t-1} \log(1 - r_\tau) = \sum_{\tau=0}^{t-1} \log\Big(1 - (\phi_{\tau-1} - \phi_\tau)\Big) \\ &\le t \log\Big(1 - \frac{1}{t}\sum_{\tau=0}^{t-1}(\phi_{\tau-1} - \phi_\tau)\Big) = t \log\Big(1 - \frac{1 - \phi_{t-1}}{t}\Big). \quad (27)\end{aligned}$$

In light of (15), ϕ_t is decreasing in t, and hence

$$1 - \prod_{\tau=0}^{t-1}(1 - r_\tau) \ge 1 - \Big(1 - \frac{1 - \phi_{t-1}}{t}\Big)^t \ge 1 - \Big(1 - \frac{1 - \phi_0}{t}\Big)^t. \quad (28)$$

To bound $1 - \Big(1 - \frac{1-\phi_0}{t}\Big)^t$ further, consider the function $f_t(x) = 1 - (1 - x/t)^t$ for x in a neighborhood of the origin, with $t \ge 1$. For any fixed x, $f_t(x)$ decreases to $1 - e^{-x}$ as t tends to infinity. The latter function is concave, and hence for all $x \le \varepsilon$,

$$f_t(x) \ge 1 - e^{-x} \ge \frac{1 - e^{-\varepsilon}}{\varepsilon} x.$$

Focusing on $1 - \phi_0$, from (18) and (17), by using Chebyshev's inequality with $\sum_{w\in W} p_w^2 = c/n$, it follows that ϕ_0 is concentrated around its mean $\mathbb{E}[\phi_0] = 1 - c/n$. Therefore, with probability $1 - o(1/n)$, $1 - \phi_0 = o(1)$. But $(1 - e^{-\varepsilon})/\varepsilon$ can be made arbitrary close to 1 by taking ε small enough, so it follows that $1 - \prod_{\tau=0}^{t-1}(1 - r_\tau) > c'/n$ for some constant $c' \in (1, c)$ arbitrarily close to c. Hence, the branching process on RIG is stochastically lower bounded by $\mathrm{Bin}(n-1, c'/n)$. But this bound itself stochastically dominates a branching process on $G_{n,c'/n}$. Because $c' > 1$, there exists **whp** a giant component of size $\Theta(n)$ in $G_{n,c'/n}$. This implies that the stopping time of the branching process associated to $G_{n,c'/n}$ is $\Theta(n)$ with high probability, as is therefore the stopping time T_v for some $v \in V$. Thus, **whp** there is a giant component in a general RIG.

We now show that this giant component is unique and that all other components have size $O(\log n)$. Consider the size of the giant component. From the representation (15) for ϕ_{t-1}, consider the previously introduced random variables $X_{t,w} = n\log(1/(1-p_w))\mathbb{I}(\Gamma_w \leq t)$. Similarly to the proof of Theorem 1, it follows that under the conditions of the theorem there is a positive constant $\delta > 0$ such that $\sum_w X_{t,w}$ is concentrated within $(1 \pm \delta)\sum_w \mathbb{E}[X_{t,w}] = (1 \pm \delta)c/n$, with probability $1 - o(1)$. Hence, there exists $p^+ = c^+/n$, for some constant $c^+ > c > 1$, such that $1 - \phi_{t-1} \leq 1 - (1-p^+)^t$, which is equivalent to $-\log \phi_{t-1} \leq t\log(1-p^+) = tp^+ + o(tp^+) = tc^+/n + o(t/n)$. Similarly, the concentration of ϕ_{t-1} implies that there exists $p^- = c^-/n$, with $c > c^- > 1$, such that $1-(1-p^-)^t \leq 1-(1-(1-\phi_{t-1})/t)^t$, which implies that $-\log\phi_{t-1} \geq t\log(1-p^-) = tp^- + o(tp^-) = tc^-/n + o(t/n)$. Combining the upper and lower bound, we conclude that, with probability $1-o(1)$, the rate of the branching process on RIG is bracketed by

$$1-(1-p^-)^t \leq 1 - \prod_{\tau=0}^{t-1}(1-r_\tau) \leq 1-(1-p^+)^t. \tag{29}$$

The stochastic dominance of the binomial distribution, together with (29), implies

$$\begin{aligned}\mathbb{P}\Big[\mathrm{Bin}\Big(n-1, 1-(1-p^-)^t\Big) \geq t\Big] &\leq \mathbb{P}\Big[\mathrm{Bin}\Big(n-1, 1-\prod_{\tau=0}^{t-1}(1-r_\tau)\Big) \geq t\Big] \\ &\leq \mathbb{P}\Big[\mathrm{Bin}\Big(n-1, 1-(1-p^+)^t\Big) \geq t\Big]. \end{aligned} \tag{30}$$

In light of (29), the branching process $\{Y_t\}_{t\geq 0}$ associated to a RIG is stochastically bounded from below and above by the branching processes associated with G_{n,p^-} and G_{n,p^+}, respectively [2]. Since both $c^-, c^+ > 1$, there exist giant components in both G_{n,p^-} and G_{n,p^+}, **whp**.

In [24], it has been shown that the giant components in $G_{n,\lambda/n}$, for $\lambda > 1$, is unique and of size $\approx n\zeta_\lambda$, where ζ_λ is the unique solution in $(0,1)$ of the equation

$$1 - e^{-\lambda\zeta} = \zeta. \tag{31}$$

Moreover, the size of the giant component in $G_{n,\lambda/n}$ satisfies the central limit theorem

$$\frac{\max_v\{|C(v)\}| - \zeta_\lambda n}{\sqrt{n}} \sim \mathcal{N}\Big(0, \frac{\zeta_\lambda(1-\zeta_\lambda)}{(1-\lambda+\lambda\zeta_\lambda)^2}\Big). \tag{32}$$

From the definition of the stopping time (see (19)) and given (30) and (32), there is a giant component in a RIG of size at least $n\zeta_\lambda(1-o(1))$, **whp**. Furthermore, the stopping times of the branching processes associated to G_{n,p^-} and G_{n,p^+} are approximately ζn, where ζ satisfies (31), with $\lambda^- = np^-$ and $\lambda^+ = np^+$, respectively. These two stopping times are close to one another, which follows from analyzing the function $F(\zeta, c) = 1 - \zeta - e^{-c\zeta}$, where (ζ, c) is the solution of $F(\zeta, c) = 0$, for given c. Since all partial derivatives of $F(\zeta, c)$ are continuous and bounded, the stopping times of the branching processes defined from G_{n,p^-}, G_{n,p^+} are "close" to the solution of (31), for

$\lambda = c$. From (30), the stopping time of a RIG is bounded by the stopping times on G_{n,p^-}, G_{n,p^+}.

For the last part of the proof of uniqueness of the giant component, we adapt the arguments in [2] to our setting. Let us assume that there are at least two giant components in a RIG, with the sets of nodes $V_1, V_2 \subset V$. Let us create a new, independent "sprinkling" $\widehat{\text{RIG}}$ on the top of our RIG, with the same sets of nodes and attributes, while $\hat{p}_w = p_w^\gamma$, for $\gamma > 1$ to be defined later. Now, our object of interest is $\text{RIG}_{new} = \text{RIG} \cup \widehat{\text{RIG}}$. Let us consider all $\Theta(n^2)$ pairs $\{v_1, v_2\}$, where $v_1 \in V_1, v_2 \in V_2$, which are independent in $\widehat{\text{RIG}}$, (but not in RIG), hence the probability that two nodes $v_1, v_2 \in V$ are connected in $\widehat{\text{RIG}}$ is given by

$$1 - \prod_w (1 - \hat{p}_w^2) = 1 - \prod_w (1 - p_w^{2\gamma}) = \sum_w p_w^{2\gamma} + o(\sum_w p_w^{2\gamma}), \tag{33}$$

since $\gamma > 1$ and $p_w = O(1/n)$ for any w. Given that $\sum_w p_w^2 = c/n$, we choose $\gamma > 1$ so that $\sum_w p_w^{2\gamma} = \omega(1/n^2)$. Now, by the Markov inequality, **whp** there is a pair $\{v_1, v_2\}$ such that v_1 is connected to v_2 in $\widehat{\text{RIG}}$, implying that V_1, V_2 are connected, **whp**, forming one connected component within RIG_{new}. From the previous analysis, it follows that this component is of size at least $2n\zeta_\lambda(1-\delta)$ for any small constant $\delta > 0$. On the other hand, the probabilities p_w^{new} in RIG_{new} satisfy

$$p_w^{new} = 1-(1-p_w)(1-\hat{p}_w) = p_w + \hat{p}_w(1-p_w) = p_w + p_w^\gamma(1-p_w) = p_w(1+o(1)),$$

again since $\gamma > 1$ and $p_w = O(1/n)$ for any w. Thus,

$$\sum_{w\in W} (p_w^{new})^2 = \sum_{w\in W} p_w^2 + \Theta(\sum_{w\in W} p_w^{1+\gamma}(1-p_w)) = \sum_{w\in W} p_w^2(1+o(1)) = c/n + o(1/n). \tag{34}$$

Given that the stopping time on RIG is bounded by the stopping times on G_{n,p^-}, G_{n,p^+}, and from its continuity, it follows that the giant component in RIG_{new} cannot be of size $2n\zeta_\lambda(1-\delta)$, which is a contradiction. Thus, there is only one giant component in RIG, of size given by $n\zeta_c(1+o(1))$, where ζ_c satisfies (31), for $\lambda = c$. Moreover, knowing the behavior of $G_{n,p}$, from (30), it follows that all other components are of size $O(\log n)$.

6 Conclusion

The analysis of random models for bipartite graphs is important for the study of algorithms on networks formed by associating nodes with shared attributes. In the random intersection graph (RIG) model, nodes have certain attributes with fixed probabilities. In this paper, we have considered the general RIG model, where these probabilities are represented by a set of probabilities $\boldsymbol{p} = \{p_w : w \in W\}$, where p_w denotes the probability that a node is attached to the attribute w.

We have analyzed the evolution of components in general RIGs, giving conditions for existence and uniqueness of the giant component. We have done so by generalizing the branching process argument used to study the birth of the giant component in Erdős-Rényi graphs. We have considered a dependent, inhomogeneous Galton-Watson

process, where the number of offspring follows a binomial distribution with a different number of nodes and different rate at each step during the evolution. The analysis of such a process is complicated by the dependence on its history, dictated by the structure of general RIGs. We have shown that in spite of this difficulty, it is possible to give stochastic bounds on the branching process, and that under certain conditions the giant component appears at the threshold $n \sum_{w \in W} p_w^2 = 1$, with probability tending to one, as the number of nodes tends to infinity.

Acknowledgments

Part of this work was funded by the Department of Energy ASCR program, by the Air Force Office of Scientific Research MURI grant FA9550-10-1-0569, and by the Office of Naval Research grant N00014-10-1-0641. Nicolas W. Hengartner was supported by DOE-LDRD 20080391ER.

References

1. ALBERT, R., AND BARABÁSI, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys. 74*, 1 (2002), 47 – 97.
2. ALON, N., AND SPENCER, J. H. *The probabilistic method*, 2nd ed. John Wiley & Sons, Inc., New York, 2000.
3. BARABÁSI, A. L., AND ALBERT, R. Emergence of Scaling in Random Networks. *Science 286*, 5439 (1999), 509–512.
4. BEHRISCH, M. Component evolution in random intersection graphs. In *Electr. J. Comb.* (2007), vol. 14.
5. BERNSTEIN, S. N. On a modification of chebyshevs inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math. 4*, 25 (1924).
6. BLOZNELIS, M., JAWORSKI, J., AND RYBARCZYK, K. Component evolution in a secure wireless sensor network. *Netw. 53*, 1 (2009), 19–26.
7. CHUNG, F., AND LU, L. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences of the United States of America 99*, 25 (2002), 15879–15882.
8. DEIJFEN, M., AND KETS, W. Random intersection graphs with tunable degree distribution and clustering. *Probab. Eng. Inf. Sci. 23*, 4 (2009), 661–674.
9. ERDŐS, P., GOODMAN, A. W., AND PÓSA, L. The representation of a graph by set intersections. *Canad. J. Math. 18* (1966), 106–112.
10. ERHARD GODEHARDT, JERZY JAWORSKI, K. R. Random intersection graphs and classification. In *Advances in Data Analysis* (2007), vol. 45, pp. 67–74.
11. EUBANK, S., GUCLU, H., ANIL KUMAR, V. S., MARATHE, M. V., SRINIVASAN, A., TOROCZKAI, Z., AND WANG, N. Modelling disease outbreaks in realistic urban social networks. *Nature 429*, 6988 (May 2004), 180–184.
12. FILL, J. A., SCHEINERMAN, E. R., AND SINGER-COHEN, K. B. Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $g(n, m, p)$ and $g(n, p)$ models. *Random Struct. Algorithms 16*, 2 (2000), 156–176.
13. GODEHARDT, E., AND JAWORSKI, J. Two models of random intersection graphs and their applications. *Electronic Notes in Discrete Mathematics 10* (2001), 129–132.
14. GUILLAUME, J.-L., AND LATAPY, M. Bipartite graphs as models of complex networks. *Physica A: Statistical and Theoretical Physics 371*, 2 (2006), 795 – 813.

15. JAWORSKI, J., AND STARK, D. The vertex degree distribution of passive random intersection graph models. *Comb. Probab. Comput. 17*, 4 (2008), 549–558.
16. KAROŃSKI, M., SCHEINERMAN, E., AND SINGER-COHEN, K. On random intersection graphs:the subgraph problem. *Combinatorics, Probability and Computing 8* (1999).
17. LAGERÅS, A. N., AND LINDHOLM, M. A note on the component structure in random intersection graphs. *Electronic Journal of Combinatorics 15*, 1 (2008).
18. NEWMAN, M. E. J. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E 64*, 1 (Jun 2001), 016131.
19. NEWMAN, M. E. J., AND PARK, J. Why social networks are different from other types of networks. *Phys. Rev. E 68*, 3 (Sep 2003), 036122.
20. NEWMAN, M. E. J., STROGATZ, S. H., AND WATTS, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E 64*, 2 (Jul 2001), 026118.
21. NIKOLETSEAS, S., RAPTOPOULOS, C., AND SPIRAKIS, P. Large independent sets in general random intersection graphs. *Theor. Comput. Sci. 406* (October 2008), 215–224.
22. NIKOLETSEAS, S. E., RAPTOPOULOS, C., AND SPIRAKIS, P. G. The existence and efficient construction of large independent sets in general random intersection graphs. In *ICALP* (2004), J. Díaz, J. Karhumäki, A. Lepistö, and D. Sannella, Eds., vol. 3142 of *Lecture Notes in Computer Science*, Springer, pp. 1029–1040.
23. NIKOLETSEAS, S. E., RAPTOPOULOS, C., AND SPIRAKIS, P. G. Expander properties and the cover time of random intersection graphs. *Theor. Comput. Sci. 410*, 50 (2009), 5261–5272.
24. REMCO VAN DER HOFSTAD. Random graphs and complex networks. Lecture notes in preparation, `http://www.win.tue.nl/~rhofstad/NotesRGCN.html`.
25. RYBARCZYK, K. Equivalence of the random intersection graph and $G(n, p)$, 2009. Submitted, `http://arxiv.org/abs/0910.5311`.
26. SINGER-COHEN, K. Random intersection graphs. PhD thesis, Johns Hopkins University, 1995.
27. WATTS, D. J., AND STROGATZ, S. H. Collective dynamics of Small-World networks. *Nature 393*, 6684 (1998), 440–442.